# Introduction to Bayesian inference

End of course data analysis project

—

The Center for Disease Control (CDC) reports the vaccination coverage of Varicella among young children. Varicella, commonly known as chicken- pox, is a highly contagious viral infection caused by the varicella-zoster virus (VZV). Vaccination against chickenpox has been highly effective in reducing the incidence and severity of the disease. In the United States, vaccination against varicella has been part of the routine childhood immunization sched- ule since the mid-1990s. Since the vaccine's introduction, there has been a dramatic decline in the number of chickenpox cases, hospitalizations, and deaths associated with the disease. The target for vaccination coverage of varicella (chickenpox) in the United States, as set by the Centers for Dis- ease Control and Prevention (CDC), is typically around 90% or higher for children. This high coverage rate is aimed at achieving herd immunity and preventing outbreaks of chickenpox within communities.

## Project 1: Insurance

The next table summarizes, based on a survey, the number of children in the birth cohort 2014-2017 that had at least one dose of the Varicella vaccine. It gives the number of vaccinated children (Vaccinated) amongst the number of children in the survey (Sample Size). The information is provided for 5 regions of the US, and split according to insurance status (private insurance, uninsured or any Medicaid).

| Geography | Insurance | Vaccinated | Sample Size |
|---|---|---|---|
| North Carolina | Any Medicaid | 380 | 419 |
| North Carolina | Private Insurance Only | 632 | 673 |
| North Carolina | Uninsured | 28 | 34 |
| Georgia | Any Medicaid | 363 | 396 |
| Georgia | Private Insurance Only | 527 | 576 |
| Georgia | Uninsured | 36 | 50 |
| Wisconsin | Any Medicaid | 282 | 332 |
| Wisconsin | Private Insurance Only | 514 | 548 |
| Wisconsin | Uninsured | 16 | 34 |
| Florida | Any Medicaid | 446 | 490 |
| Florida | Private Insurance Only | 588 | 628 |
| Florida | Uninsured | 28 | 39 |
| Mississippi | Private Insurance Only | 400 | 441 |
| Mississippi | Uninsured | 27 | 32 |

## Question 1

> Derive analytically the posterior of the vaccination coverage per ge- ography and insurance group. Use a conjugate prior that (1) reflects no knowledge on the vaccination coverage, and (2) reflects that vac- cination coverage is typically around 90% or higher. Give posterior summary measures of the vaccination coverage per geography and in- surance group. Is the choice of the prior impacting your results?

### Theoretical considerations

The outcome *Vaccinated/Not Vaccinated* follows a Bernouilli distribution with parameter $p$:

$V$ : Vaccination status $V \in \{0, 1\}$ $V \sim \mathcal{B}ern(p)$

It is known from theory that the sum of $n$ *i.i.d* Bernoulli random variables follows a Binomial distribution. This will be used to model the sample outcome: the number of vaccinated people $V_s$ in a random sample of size $n$:

$$V_s = \sum_{i}^{n} V_i \sim \mathcal{B}inom(n, \theta)$$

where $\theta$ is the parameter of interest - the vaccine coverage.

In the course, we saw that the Beta distribution is the conjugate prior for binomially distributed data:

| Distribution | Formula |
|---|---|
| Prior | $p(\theta) = \mathcal{B}eta(\alpha, \beta)$ |
| Likelihood | $p(y \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ |
| Posterior | $p(\theta \mid y) = \mathcal{B}eta(\alpha + k, \beta + n - k)$ |

The summary measures for the Beta distribution are defined as follows:

| Summary Measure | Formula |
|---|---|
| Mean | $\frac{\alpha}{\alpha+\beta}$ |
| Median | See Note |
| Mode | $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$ |

Note: The median of the Beta distribution does not have a simple closed form expression. It can be approximated numerically or using statistical software.

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 14
   Unobserved stochastic nodes: 3
   Total graph size: 86

Initializing model
```

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 14
   Unobserved stochastic nodes: 7
   Total graph size: 222

Initializing model
```

```
[1] "FL_Medicaid"
[1] "FL_Private"
[1] "FL_Uninsured"
[1] "MS_Private"
[1] "MS_Uninsured"
[1] "GA_Medicaid"
[1] "GA_Private"
[1] "GA_Uninsured"
[1] "WI_Medicaid"
[1] "WI_Private"
[1] "WI_Uninsured"
[1] "chain"
```

## Choice of prior distributions

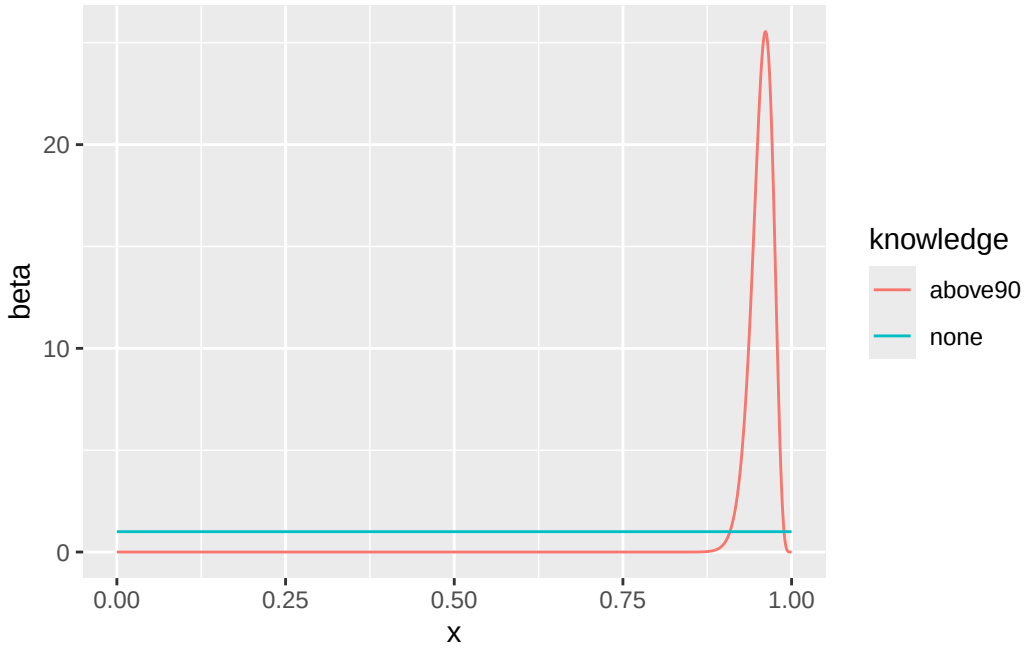
### (1) No prior knowledge

In order to reflect no prior knowledge on the vaccine coverage, the weakly-informative prior Beta(1,1) will be used, which is equivalent to the uniform distribution over $[0, 1]$.

### (2) Vaccine coverage >90%

For modeling prior knowledge that vaccine coverage is about 90%, we chose the Beta(150, 7) distribution.

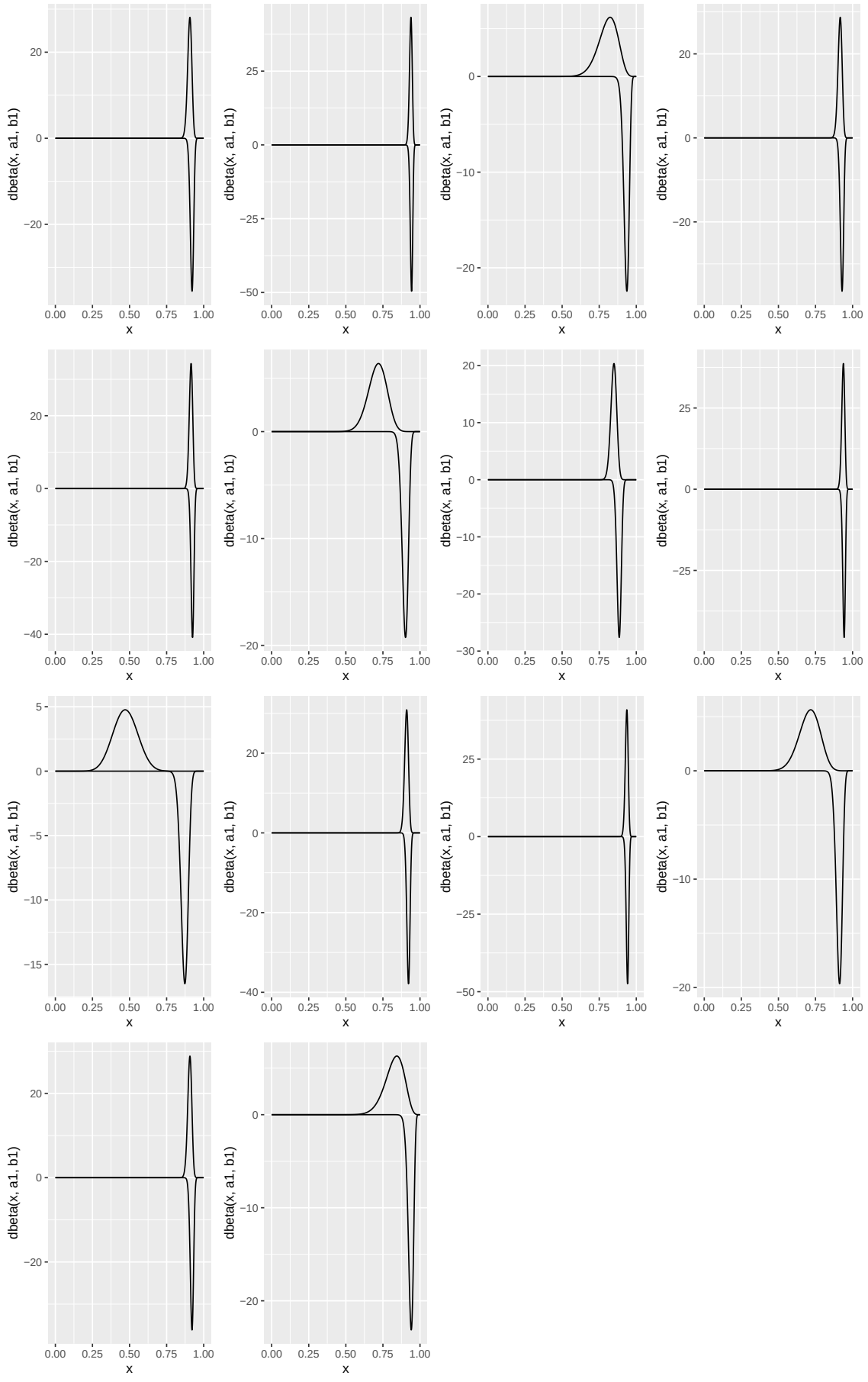
### Comparison of priors




## Results

Table 4: Posterior distribution parameters and summary measures per geography for the two different Beta priors

| | | Posterior parameters and summary measures | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Beta(1,1) prior | | | | | | | Beta(150,7) prior | | | | | | |
| Geography | Insurance | alpha | beta | mean | mode | var | HPD LL | HPD UL | alpha | beta | mean | mode | var | HPD LL | HPD UL |
| NC | Medicaid | 381 | 40 | 0.905 | 0.907 | 0.014 | 0.875 | 0.931 | 530 | 46 | 0.920 | 0.922 | 0.011 | 0.897 | 0.941 |
| NC | Private | 633 | 42 | 0.938 | 0.939 | 0.009 | 0.918 | 0.955 | 782 | 48 | 0.942 | 0.943 | 0.008 | 0.925 | 0.957 |
| NC | Uninsured | 29 | 7 | 0.806 | 0.824 | 0.065 | 0.664 | 0.916 | 178 | 13 | 0.932 | 0.937 | 0.018 | 0.892 | 0.963 |
| GA | Medicaid | 364 | 34 | 0.915 | 0.917 | 0.014 | 0.885 | 0.940 | 513 | 40 | 0.928 | 0.929 | 0.011 | 0.905 | 0.948 |
| GA | Private | 528 | 50 | 0.913 | 0.915 | 0.012 | 0.889 | 0.935 | 677 | 56 | 0.924 | 0.925 | 0.010 | 0.903 | 0.942 |
| GA | Uninsured | 37 | 15 | 0.712 | 0.720 | 0.062 | 0.583 | 0.825 | 186 | 21 | 0.899 | 0.902 | 0.021 | 0.854 | 0.936 |
| WI | Medicaid | 283 | 51 | 0.847 | 0.849 | 0.020 | 0.807 | 0.884 | 432 | 57 | 0.883 | 0.885 | 0.014 | 0.854 | 0.910 |
| WI | Private | 515 | 35 | 0.936 | 0.938 | 0.010 | 0.915 | 0.955 | 664 | 41 | 0.942 | 0.943 | 0.009 | 0.923 | 0.958 |
| WI | Uninsured | 17 | 19 | 0.472 | 0.471 | 0.082 | 0.314 | 0.634 | 166 | 25 | 0.869 | 0.873 | 0.024 | 0.818 | 0.913 |
| FL | Medicaid | 447 | 45 | 0.909 | 0.910 | 0.013 | 0.882 | 0.932 | 596 | 51 | 0.921 | 0.922 | 0.011 | 0.899 | 0.941 |
| FL | Private | 589 | 41 | 0.935 | 0.936 | 0.010 | 0.914 | 0.953 | 738 | 47 | 0.940 | 0.941 | 0.008 | 0.923 | 0.956 |
| FL | Uninsured | 29 | 12 | 0.707 | 0.718 | 0.070 | 0.561 | 0.834 | 178 | 18 | 0.908 | 0.912 | 0.021 | 0.864 | 0.944 |
| MS | Private | 401 | 42 | 0.905 | 0.907 | 0.014 | 0.876 | 0.931 | 550 | 48 | 0.920 | 0.921 | 0.011 | 0.897 | 0.940 |
| MS | Uninsured | 28 | 6 | 0.824 | 0.844 | 0.064 | 0.681 | 0.930 | 177 | 12 | 0.937 | 0.941 | 0.018 | 0.898 | 0.967 |


dbeta(x, a1, b1)
x
0.00
0.25
0.50
0.75
1.00

# Question 2

> nvestigate whether the vaccination coverage is associated with the in- surance status using a logistic regression model Yij Binom( ij ,Nij ) with logit( ij ) = 0 + 1IAnyMedicaid,ij + 2IUninsured,ij where i is the location, j is the insurance status, ij is the vaccination coverage and I. are dummy variables. Assume non-informative priors for the parameters to be estimated. Write and explain the code in BUGS language

```
model
{
  for (t in 1:T) {
    # Likelihood
    y[t] ~ dbin(p[t], K[t])
    # conditional mean model using link function
    logit(p[t]) <- alpha_0 + ins_1 * x_1[t] + ins_2 * x_2[t]
 }

  # Priors
  alpha_0 ~ dnorm(0.0,0.01)
  ins_1 ~ dnorm(0.0,0.01)
  ins_2 ~ dnorm(0.0,0.01)


  # Vaccine coverage per Insurance group
  pi_private <- exp(alpha_0)/(1+exp(alpha_0))
  pi_medicaid <- exp(alpha_0 + ins_1)/(1+exp(alpha_0 + ins_1))
  pi_uninsured <- exp(alpha_0 + ins_2)/(1+exp(alpha_0 + ins_2))

  # Difference between vaccine coverage per insurance group
  diff_priv_medicaid  <- pi_private - pi_medicaid
  diff_priv_uninsured <- pi_private - pi_uninsured
}
```

**Likelihood function and model specification**

For each row `t` in the dataset, the outcome `y[t]` (number of vaccinated children in the sample) is specified as drawn from a Binomial distribution with parameters `p[t]` and sample size `K[t]`. We also specify the conditional mean model for this likelihood function using the `logit` link function.

**Prior distribution**

As prior distribution for all parameters (intercept and indicator variables for the insurance group), a vague prior is chosen : $\mathcal{N}(\mu = 0, \tau = 0.01)$.

**Quantities of interest** During each MCMC run, the vaccine coverage per insurance group is calculated using the inverse logit transformation, this will give access to the posterior distribution of vaccine coverage per insurance group. Furthermore, the differences between vaccine coverages are calculated which will be needed to answer Question 6.

# Question 3

> Run the MCMC method and check convergence of the MCMC chains. Give the details on how you checked convergence.

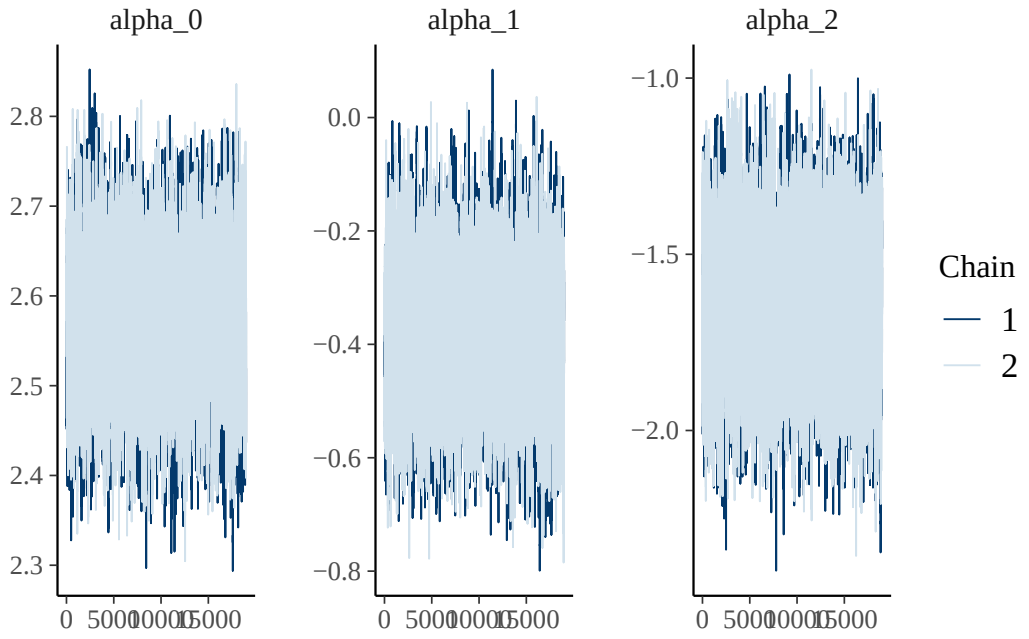
## MCMC run summary

```
Inference for Bugs model at "4", fit using jags,
 2 chains, each with 40000 iterations (first 2000 discarded), n.thin = 2
 n.sims = 38000 iterations saved. Running time = 1.644 secs
                    mu.vect sd.vect   2.5%     25%     50%     75%   97.5%
```

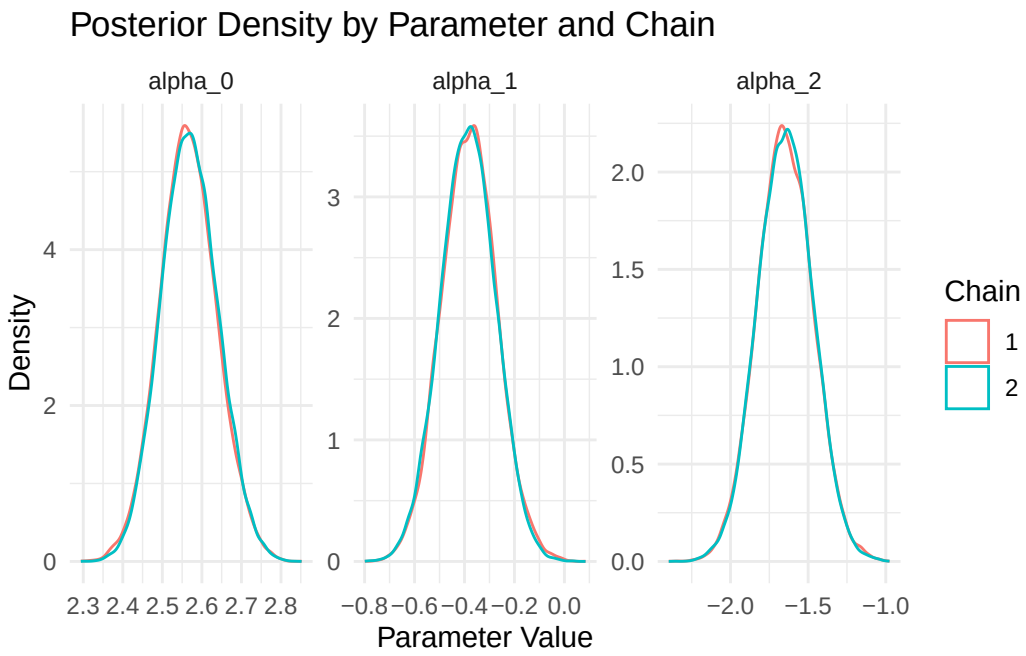
```
alpha_0               2.568   0.072  2.429   2.519   2.566   2.615   2.712
alpha_1              -0.383   0.110 -0.598  -0.458  -0.383  -0.309  -0.166
alpha_2              -1.647   0.177 -1.991  -1.767  -1.648  -1.527  -1.299
diff_priv_medicaid    0.030   0.009  0.013   0.024   0.030   0.036   0.048
diff_priv_uninsured   0.215   0.033  0.152   0.192   0.214   0.237   0.282
pi_medicaid           0.899   0.007  0.883   0.894   0.899   0.904   0.913
pi_private            0.929   0.005  0.919   0.925   0.929   0.932   0.938
pi_uninsured          0.714   0.033  0.647   0.692   0.715   0.737   0.776
deviance            102.273   2.443 99.490 100.499 101.634 103.370 108.654
                     Rhat n.eff
alpha_0             1.002  1700
alpha_1             1.002  2300
alpha_2             1.001 38000
diff_priv_medicaid  1.002  2800
diff_priv_uninsured 1.001 11000
pi_medicaid         1.001 20000
pi_private          1.002  1700
pi_uninsured        1.001  5200
deviance            1.001 38000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule: pV = var(deviance)/2)
pV = 3.0 and DIC = 105.3
DIC is an estimate of expected predictive error (lower deviance is better).
```

## Convergence checks

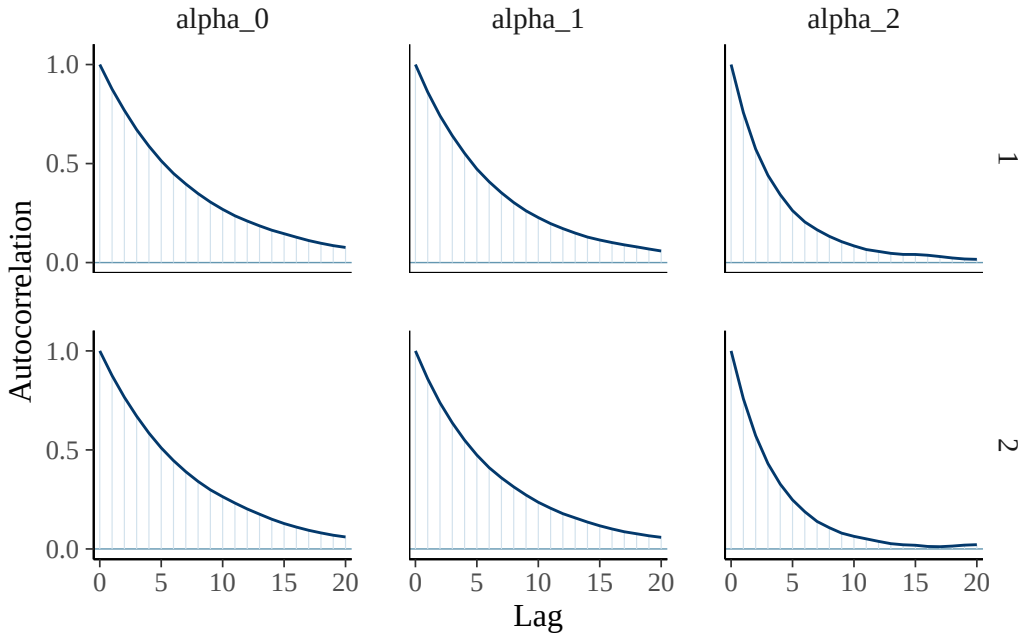
### Traceplots




### Density plots per chain




**Autocorrelation plot**




$\hat{R}$

```
Potential scale reduction factors:

        Point est. Upper C.I.
alpha_0          1       1.01
alpha_1          1       1.01
alpha_2          1       1.00

Multivariate psrf

1
```

**Geweke diagnostics**

```
[[1]]
```

```
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

alpha_0 alpha_1 alpha_2
-1.6747  0.7104  0.6236


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

alpha_0 alpha_1 alpha_2
 0.9378 -0.5204 -0.4942
```

## Question 4

> Make a plot of the posterior of the model parameters and give posterior summary measures. Interpret the results.

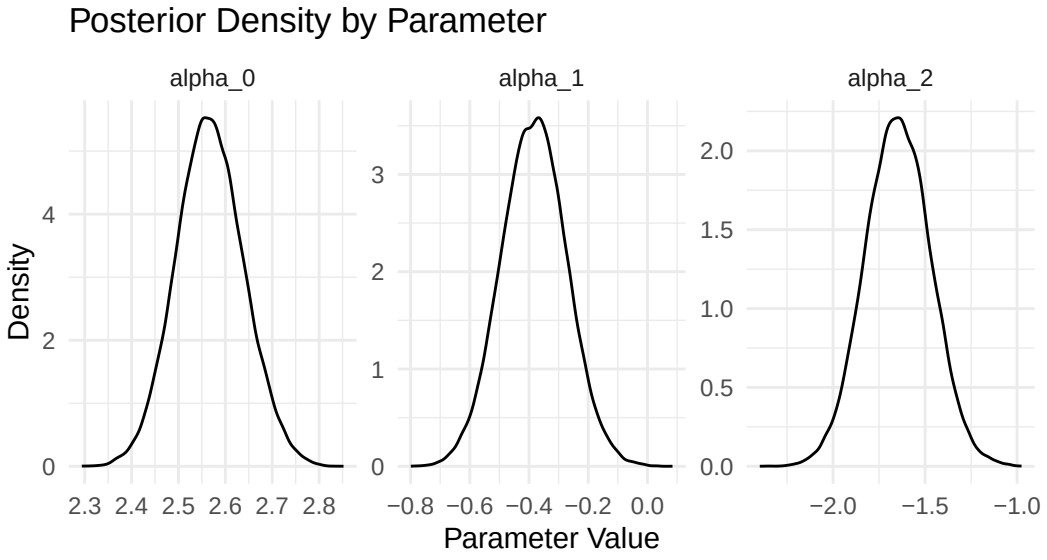



Figure 1

Table 5

| Parameter | Summary measures | | | 95% HPD interval | |
|---|---|---|---|---|---|
| | Mean | Median | SD | LL | UL |
| alpha_0 | 2.568 | 2.566 | 0.072 | 2.424 | 2.706 |
| alpha_1 | -0.383 | -0.383 | 0.110 | -0.595 | -0.163 |
| alpha_2 | -1.647 | -1.648 | 0.177 | -1.990 | -1.299 |

Plots of posterior densities and summary measures of the model parameters are given in Figure 1 and Table 5, respectively.

With `Private Insurance Only` as reference category the interpretation of the parameters is as follows:

- $\alpha_0$ gives the $log(odds)$ of Vaccinated *vs.* Non-Vaccinated in the private insurance group
- $\alpha_1$ gives the change in $log(odds)$ in the Medicaid group *vs.* the private Insurance group, the $log(odds)$ in the Medicaid group are given by $\alpha_0 + \alpha_1$

- $\alpha_2$ gives the change in $log(odds)$ in the Uninsured group *vs.* the private Insurance group, the $log(odds)$ in the Uninsured group are given by $\alpha_0 + \alpha_2$

Numerically, the posterior mean of $\alpha_0$ is estimated at 2.568 and with a posterior probability of 95% $\alpha_0$ lies in $[2.424, 2.706]$. The posterior estimates for $\alpha_1$ and $\alpha_2$ can be read in the same manner from Table 5.

# Question 5

> Give the posterior estimate of the vaccination coverage per region and insurance status. Compare with the analytical results you obtained in Question 1.

The vaccine coverage in a given group can be obtained from the $log(odds)$ (see model code Question 3):

$$\pi = \frac{\exp(log(odds))}{1 + \exp(log(odds))}$$

Posterior estimates for the mean vaccine coverage per region and insurance status are given in Table 6 and Figure 2.

Table 6: Posterior estimates of mean vaccine coverages from the logistic regression model vs. estimates from conjugate pair modeling. Subscripts indicate whether the esimates was obtained from logistic regression or the chosen prior distribution, respectively. $\Delta_{Beta(\alpha,\beta)}$ gives the difference between posterior estimates from the logistic regression model and conjugate pair modeling. Vaccine coverage for MS/Medicaid can only be obtained from the logistic regression model.

| Geo. | Ins. | $k$ | $n$ | $\bar{\pi}_{logreg}$ | $\bar{\pi}_{Beta(1,1)}$ | $\bar{\pi}_{Beta(150,7)}$ | $\Delta_{Beta(1,1)}$ | $\Delta_{Beta(150,7)}$ |
|---|---|---|---|---|---|---|---|---|
| FL | Medicaid | 446 | 490 | 0.899 | 0.909 | 0.921 | -0.010 | -0.022 |
| FL | Private | 588 | 628 | 0.929 | 0.935 | 0.940 | -0.006 | -0.011 |
| FL | Uninsured | 28 | 39 | 0.714 | 0.707 | 0.908 | 0.007 | -0.194 |
| GA | Medicaid | 363 | 396 | 0.899 | 0.915 | 0.928 | -0.016 | -0.029 |
| GA | Private | 527 | 576 | 0.929 | 0.913 | 0.924 | 0.016 | 0.005 |
| GA | Uninsured | 36 | 50 | 0.714 | 0.712 | 0.899 | 0.002 | -0.185 |
| MS | Private | 400 | 441 | 0.929 | 0.905 | 0.920 | 0.024 | 0.009 |
| MS | Uninsured | 27 | 32 | 0.714 | 0.824 | 0.937 | -0.110 | -0.223 |
| MS | Medicaid | NA | NA | 0.899 | NA | NA | NA | NA |
| NC | Medicaid | 380 | 419 | 0.899 | 0.905 | 0.920 | -0.006 | -0.021 |
| NC | Private | 632 | 673 | 0.929 | 0.938 | 0.942 | -0.009 | -0.013 |
| NC | Uninsured | 28 | 34 | 0.714 | 0.806 | 0.932 | -0.092 | -0.218 |
| WI | Medicaid | 282 | 332 | 0.899 | 0.847 | 0.883 | 0.052 | 0.016 |
| WI | Private | 514 | 548 | 0.929 | 0.936 | 0.942 | -0.007 | -0.013 |
| WI | Uninsured | 16 | 34 | 0.714 | 0.472 | 0.869 | 0.242 | -0.155 |

# Question 6

> Based on the logistic regression model, what is the probability (a posteriori) that coverage amongst children that have private insurance is higher than amongst children that have any medicaid? And compared to children with no insurance?

To answer this question, the differences $\pi_{private} - \pi_{medicaid}$ and $\pi_{private} - \pi_{uninsured}$ were incorporated and observed during the MCMC run of the model specified in Question 3. Posterior densities and empirical CDF are shown in Figure 3.

The probabilities of interest defined below and can be approximated by using the posterior samples from the MCMC runs :

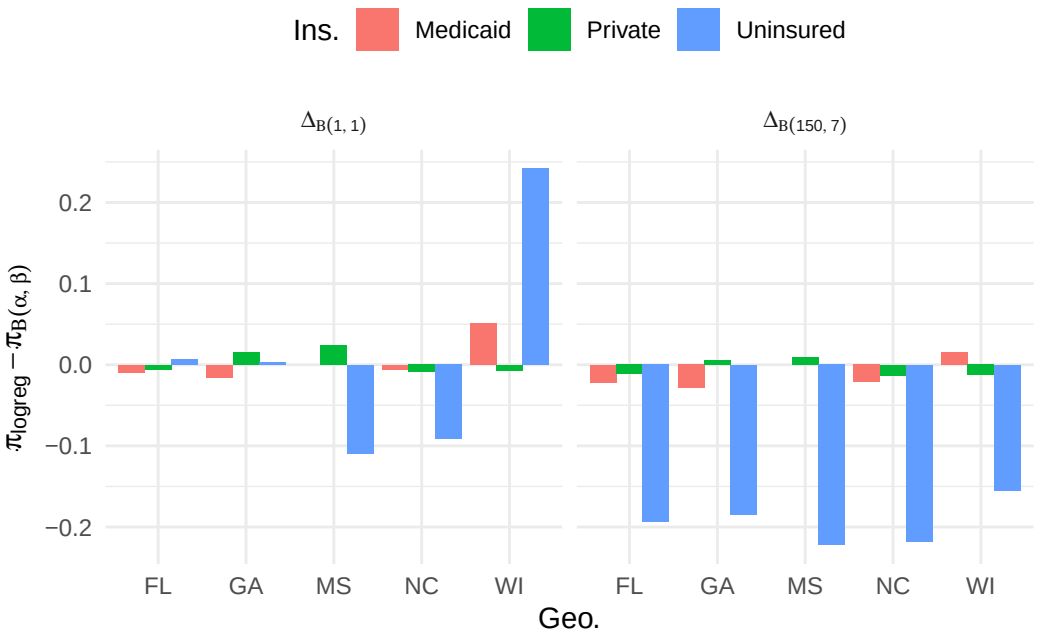



Figure 2: Differences between posterior means of vaccine coverage per region and insurance status obtained from logistic regression and conjugate pair modeling. Differences are most pronounced for the Uninsured group in MS, NC, WC when using a $Beta(1,1)prior$ and all states when using a $Beta(150,7)$ prior.

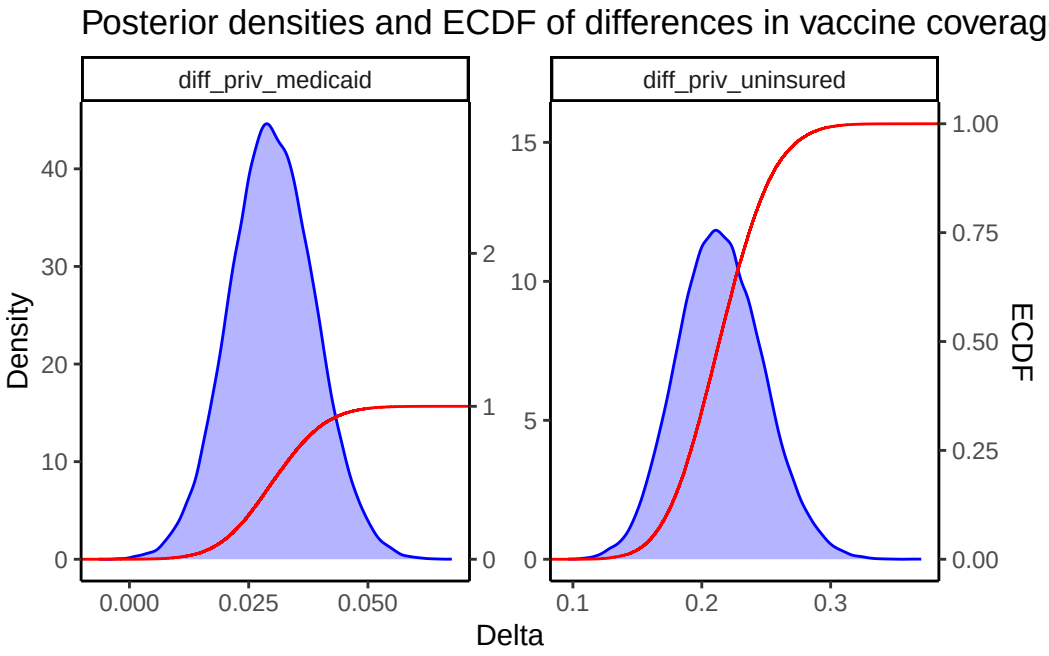



Figure 3

1. $P(\pi_{priv} > \pi_{medicaid}) = P(\pi_{priv} - \pi_{medicaid} > 0) = 0.99979$
2. $P(\pi_{priv} > \pi_{uninsured}) = P(\pi_{priv} - \pi_{uninsured} > 0) = 1$

# Question 7

> Secondly, investigate whether the vaccination coverages are distinct at the different locations by adding a location-specific intercept.
>
> $$\text{logit}\left(\pi_{ij}\right) = \alpha_{0i} + \alpha_1 I_{\text{AnyMedicaid}} + \alpha_2 I_{\text{Uninsured}}$$
>
> Assume non-informative priors for the parameters to be estimated. Write the code in BUGS language. Give a brief summary of the convergence checks you performed. Compare posteriors of vaccination coverages with results from Question 1.

```
model
{
  # Likelihood
  for (t in 1:T) {
    y[t] ~ dbin(p[t], K[t])
    logit(p[t]) <- inprod(beta, X[t,])
 }

  # Priors
  for (i in 1:7){
    beta[i] ~ dnorm(0.0,0.01)
  }

  # Vaccine coverage
  for (t in 1:T){
    #logodds
    lo[t] <- inprod(beta, X[t,])
    #convert to probability / vaccine coverage
    coverage[t] <- exp(lo[t]) / (1 + exp(lo[t]))
  }
}
```

Convergence was checked in a similar way as for the model defined in Question 3. In brief, trace plots showed the expected caterpillar pattern, posterior densities per parameter and chain showed superposed well and autocorrelation plots revealed decreasing autocorrelation with increasing lag number - all indicating convergence.

Table 7: Posterior estimates of mean vaccine coverages from the logistic regression model including a region specific intercept vs. estimates from conjugate pair modeling. Subscripts indicate whether the esimates was obtained from logistic regression or the chosen prior distribution, respectively. $\Delta_{Beta(\alpha,\beta)}$ gives the difference between posterior estimates from the logistic regression model and conjugate pair modeling. Vaccine coverage for MS/Medicaid can only be obtained from the logistic regression model.

| Geo. | Ins. | $k$ | $n$ | $\bar{\pi}_{logreg}$ | $\bar{\pi}_{Beta(1,1)}$ | $\bar{\pi}_{Beta(150,7)}$ | $\Delta_{Beta(1,1)}$ | $\Delta_{Beta(150,7)}$ |
|---|---|---|---|---|---|---|---|---|
| FL | Medicaid | 446 | 490 | 0.907 | 0.909 | 0.921 | -0.001 | -0.014 |
| FL | Private | 588 | 628 | 0.937 | 0.935 | 0.940 | 0.002 | -0.003 |
| FL | Uninsured | 28 | 39 | 0.741 | 0.707 | 0.908 | 0.034 | -0.167 |
| GA | Medicaid | 363 | 396 | 0.896 | 0.915 | 0.928 | -0.018 | -0.031 |
| GA | Private | 527 | 576 | 0.929 | 0.913 | 0.924 | 0.016 | 0.006 |
| GA | Uninsured | 36 | 50 | 0.718 | 0.712 | 0.899 | 0.006 | -0.181 |
| MS | Private | 400 | 441 | 0.918 | 0.905 | 0.920 | 0.013 | -0.002 |
| MS | Uninsured | 27 | 32 | 0.685 | 0.824 | 0.937 | -0.138 | -0.251 |
| MS | Medicaid | NA | NA | 0.880 | NA | NA | NA | NA |
| NC | Medicaid | 380 | 419 | 0.911 | 0.905 | 0.920 | 0.006 | -0.009 |

| Geo. | Ins. | $k$ | $n$ | $\bar{\pi}_{logreg}$ | $\bar{\pi}_{Beta(1,1)}$ | $\bar{\pi}_{Beta(150,7)}$ | $\Delta_{Beta(1,1)}$ | $\Delta_{Beta(150,7)}$ |
|---|---|---|---|---|---|---|---|---|
| NC | Private | 632 | 673 | 0.940 | 0.938 | 0.942 | 0.002 | -0.002 |
| NC | Uninsured | 28 | 34 | 0.751 | 0.806 | 0.932 | -0.054 | -0.181 |
| WI | Medicaid | 282 | 332 | 0.872 | 0.847 | 0.883 | 0.025 | -0.011 |
| WI | Private | 514 | 548 | 0.912 | 0.936 | 0.942 | -0.024 | -0.030 |
| WI | Uninsured | 16 | 34 | 0.667 | 0.472 | 0.869 | 0.195 | -0.202 |

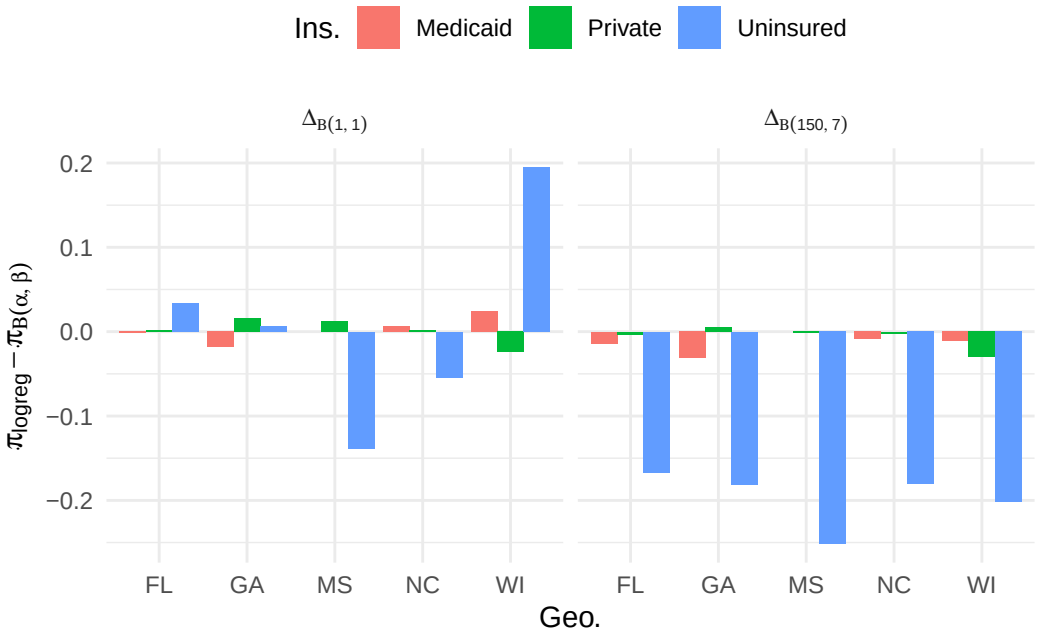



Figure 4

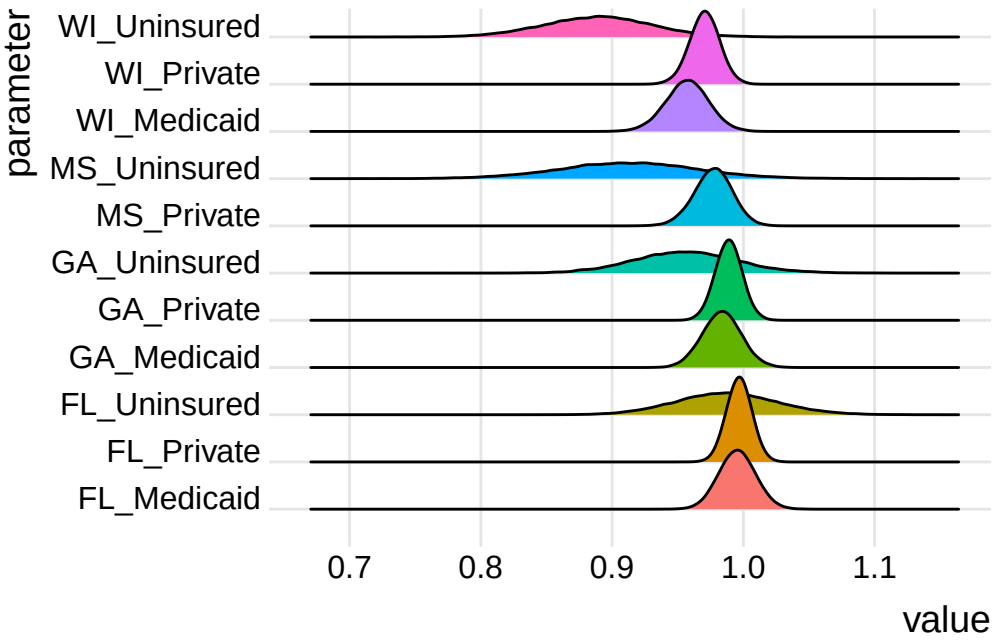
# Question 8

Compare the vaccination coverage in each of the location with the vaccination coverage in North Carolina:

$$\theta_{ij} = \frac{\pi_{ij}}{\pi_{\text{North Carolina },j}}$$

Interpret the results.

| | lower | upper |
|---|---|---|
| FL_Medicaid | 0.9683710 | 1.0229661 |
| FL_Private | 0.9782497 | 1.0153829 |
| FL_Uninsured | 0.9100095 | 1.0625293 |
| MS_Private | 0.9488678 | 1.0058223 |
| MS_Uninsured | 0.8054099 | 1.0174217 |
| GA_Medicaid | 0.9549517 | 1.0128663 |
| GA_Private | 0.9693647 | 1.0091692 |
| GA_Uninsured | 0.8742395 | 1.0320794 |
| WI_Medicaid | 0.9246291 | 0.9892471 |
| WI_Private | 0.9480345 | 0.9927000 |
| WI_Uninsured | 0.8041088 | 0.9690978 |




## Question 9

> Make a caterpillar plot of the estimated coverage (per location and insurance status). Include also the observed vaccination proportion in the plot.

## Question 10

> No data is given for children insured by Any medicaid in Mississippi. Predict the number of vaccinated individuals in Mississippi among 519 children with any medicaid